DOCUMENT RESUME

ED 346 154 TM 018 497

AUTHOR Chou, Tungshan; Wang, Lih-Shing
TITLE Making Simultaneous Inferences Using Johnson-Neyman Technique.
PUB DATE Apr 92
NOTE 16p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Computer Simulation; Equations (Mathematics); Foreign Countries; *Hypothesis Testing; *Mathematical Models; Monte Carlo Methods; *Regression (Statistics); *Sample Size; *Statistical Inference
IDENTIFIERS *Johnson Neyman Technique; Null Hypothesis; *Slope Homogeneity Test

ABSTRACT

P. O. Johnson and J. Neyman (1936) proposed a general linear hypothesis testing procedure for testing the null hypothesis of no treatment difference in the presence of some covariates. This is generally known as the Johnson-Neyman (JN) technique. The need for the hypothesis testing step (often omitted) as originally presented and the appropriateness of making simultaneous inferences after the slope homogeneity assumption test were investigated. Three regression settings were used to simulate the conditions of slight, moderate, and severe slope heterogeneity. Within each setting, 3 sample size ratios (10:10, 20:20, and 30:30, respectively) were considered with 10,000 simulated experiments in each sample size ratio. Within 9 artificially generated data conditions, the total number of simulated experiments was 90,000. Simulation results indicate that the hypothesis testing procedure as originally presented was unnecessary, whereas the slope homogeneity test was important for making simultaneous inference. When the slope homogeneity test was rejected, the simultaneous error rate was found to approximate the nominal alpha level as set forth prior to conducting the research. A caution is issued against applying the JN technique when sample sizes are small. Seven tables present analysis results, and there is a nine-item list of references. (SLD)

Making Simultaneous Inferences Using Johnson-Neyman Technique

Tungshan Chou Lih-Shing Wang

National Hualien Teachers College

Paper Presented at the Annual Meeting of the
American Educational Research Association, San Francisco, April, 1992

ABSTRACT

This paper reviewed a long-forgotten aspect of the Johnson-Neyman (J-N) technique: hypothesis testing. The originally proposed J-N technique was a two-step procedure: (1) hypothesis testing -- test whether there is a treatment effect somewhere in the entire covariate score range; if the answer is yes,we then proceed to (2) find the region of significance. Instead of performing the first step, educational researchers usually do a test of the slope homogeneity assumption. If the slope homogeneity assumption is rejected, the region of significance is then computed. It has been shown by some researchers that the region of significance as derived by Johnson and Neyman was non-simultaneous. Nevertheless, educational researchers typically make simultaneous inferences based on the computed region of significance. The purpose of this study was to investigate the need for the hypothesis testing step as originally presented in Johnson and Neyman's paper, and the appropriateness of making simultaneous inference after the slope homogeneity assumption test.

Three regression settings were employed to simulate the conditions of slight, moderate, and severe extent of slope heterogeneity. Within each setting, three sample size ratios were considered (10:10, 20:20, and 30:30) with 10,000 simulated experiments in each sample size ratio. Within nine artificially generated data conditions, the total number of simulated experiments in this study was 90,000. The simulation results indicated that the hypothesis testing procedure as originally presented was unnecessary, whereas the slope homogeneity test commonly performed before the application of the J-N technique was important for making simultaneous inference. When the slope homogeneity test was rejected, the simultaneous error rate was found to approximate the nominal alpha level as set forth by the researcher prior to conducting the research. Nevertheless, a caution was issued against the application of the J-N technique when sample sizes are small.

INTRODUCTION

Johnson and Neyman (1936) proposed a general linear hypothesis testing procedure for testing the null hypothesis of no treatment difference in the presence of some covariates. This is generally known as the Johnson-Neyman (J-N) technique of which ANCOVA and gain score analysis are only two special cases.

The original J-N technique was presented in the context of two treatment groups (of sizes n_1 and n_2, respectively) and two covariates. For the sake of simplicity and without any loss in generalizability, only one covariate is considered in this paper. Let Y be the criterion variable, X be the covariate. Johnson and Neyman expressed the expected value of the criterion variable as a function of X, i.e.,

$$E(Y) = F_1(X) = a_0 + a_1X \text{ for group 1, and}$$
$$E(Y) = F_2(X) = b_0 + b_1X \text{ for group 2.} \qquad [1]$$

Linear Hypothesis

The hypothesis, H(X), tested was that two treatment group means are equal. Unlike the two-group t-test, the hypothesis posted here takes into consideration the concommitant variable system. It should also be noted that the H(X) was not intended for any particular system of fixed values; say comparing the treatment group mean difference only at $X = X_1$ or $X = X_2$. Instead, the research question of interest was "Are there a system of values of X for which the hypothesis H(X) should be rejected?" Therefore the null hypothesis tested by the J-N technique was "There is no system of values of X at which two treatment means are different (see Johnson & Fay, 1950, p. 351)." This null hypothesis was expressed in the following linear form,

$$H(X) : a_0 - b_0 + (a_1 - b_1)X = 0. \qquad [2]$$

Test Criterion

The test statistic for the above hypothesis involved the computation of the likelihood criterion L (Johnson and Neyman, 1936, p. 77),

$$L = \frac{SSE_a}{SSE_r}, \qquad [3]$$

where SSE_a is the absolute minimum error sum of squares from fitting all four regression parameters as in expression [1] for two groups separately (SSE_a is the sum of two error sum of squares for two groups), and SSE_r is the relative error sum of squares from imposing the restrictions as set forth in the null hypothesis. Let Y_{ij} denote the outcome score for ith individual in group j, and n_j denote the number of observations in group j. The SSE_a term is obtained as

$$SSE_a = \sum_{i=1}^{n_1} (Y_{i1} - a_0 - a_1X_i)^2 + \sum_{i=1}^{n_2} (Y_{i2} - b_0 - b_1X_i)^2, \qquad [4]$$

and

$$SSE_r = \sum_{i=1}^{n_1} (Y_{i1} - a_0 - a_1X_i)^2 + \sum_{i=1}^{n_2} (Y_{i2} - a_0 - a_1X_i)^2. \quad [5]$$

Since SSE_a cannot be greater than SSE_r, L is bounded by 0 and 1. The smaller L, the less likely the null hypothesis is to be true. The distribution of L assumes the form of a Beta probability distribution with two parameters as $p = \frac{1}{2}(n_1 + n_2 - s)$ and $q = \frac{1}{2}r$. The value of s is the number of independent parameters of which the population mean is assumed to be a function with known coefficients (s = 4 as in expression [1]) and r is the number of equations required to express the hypothesis tested (r = 2 for $a_0 = b_0$, $a_1 = b_1$). A table of the values of L at various significance levels can be found in Tables of the Incomplete Beta Function (Pearson, 1956). Johnson and Neyman (1936) also presented a simplified Incomplete Beta Function table for significance levels .01 and .05, at some values of p and q.

Relationship Between L and Snedecor's F Distribution

It is seen that L is a ratio of one sum of squares to two sums of squares, expressed in terms of χ^2 as

$$L = \frac{\chi^2_a}{\chi^2_a + \chi^2_1}, \quad [6]$$

where χ^2_1 is the "extra" component due to chance fluctuations with degrees of freedom of r. The χ^2_a has $(n_1 + n_2 - s)$ degrees of freedom. Bickel and Doksum (1977, p. 13-17) discussed the relationship between the Incomplete Beta Function and the Snedecor's F distribution

$$F_{(v,\ r)} = \frac{L}{1 - L}\left(\frac{r}{v}\right), \quad [7]$$

where v is the degrees of freedom associated with SSE_a. Taking the inverse of the value obtained in equation [7], it becomes the familiar central F distribution with degrees of freedom r and v. Therefore, the p-value of the test statistic in [3] can also be obtained from an F distribution as obtained from [7].

An Easy Way to Obtain $F_{(r,v)}$

Define a dummy variable T such that $T_i = 1$ when the observation is from group one, and $T_i = 0$ otherwise. Combining the two regression lines in expression [1], we may reparameterize the model as

$$E(Y) = \beta_o + \beta_1 T + \beta_2 X + \beta_3 TX, \quad [8]$$

where $\beta_1 = (a_0 - b_0)$ and $\beta_3 = (a_1 - b_1)$. The null hypothesis H(X) can be expressed as

$$H(X):\ \beta_1 = 0,\ \beta_3 = 0.$$

The linear model under null hypothesis is

$$E(Y) = \beta_o + \beta_2 X \quad [9]$$

The F value in expression [7] can simply be obtained as follows:

$$F_{(r,v)} = \frac{(SSE_9 - SSE_8)/r}{SSE_8/(N - s)}$$

where SSE_9 is the error sum of squares associated with expression [9], SSE_8 is the error sum of squares associated with expression [8], $N = n_1 + n_2$, r and s are defined as before.

Role of Hypothesis Testing in Johnson-Neyman Technique

The first step of the originally proposed J-N technique was to perform the omnibus hypothesis H(X) testing procedure as described above. If H(X) is rejected, we conclude that the treatment difference exists over a set of the covariate points, which have been referred to as the "region of significance", denoted by R. The computation of the region of significance can be found in many statistical methods books (Huitema, 1980; Pedhazur, 1982) If we fail to reject H(X), the problem is complete at this point, and no attempt should be made to compute the region of significance.

Non-Simultaneous Versus Simultaneous Inferences

Using the scheme of comparing two regression lines, Potthoff (1964) and Rogosa (1980) rightfully pointed out that the derived region of significance as originally proposed by Johnson and Neyman (1936) is non-simultaneous. The treatment difference can be validly inferred only for any single covariate point over the region of significance, not for all points over R simultaneously. For most educational researchers however, the purpose of using the J-N technique is to find a set of the X values such that one may claim that the treatment difference exists for all X points over R at a prespecified α level. Furthermore, the covariates used in educational and psychological research are mostly random, non-simultaneous inference is seldom meaningful. Since the exact error probability of making a simultaneous inference error based on the non-simultaneous region of significance when the covariate is random can not be theoretically derived, the extent of the inappropriateness of making such simultaneous inferences is unknown.

THE CURRENT STUDY

Potthoff (1964) applied the Scheffé-like procedure to the original J-N technique to derive a simultaneous region of significance. Nevertheless, this procedure has rarely been adopted by educational researchers, perhaps due to its infereior statistical power (Chou and Huberty, 1992). Educational researchers often make simultaneous inferences based on R as yielded by the original J-N technique. Hence, the purpose of this paper is to examine the empirical performance of the J-N technique with respect to the appropriateness of making simultaneous inferences under various simulated data settings.

Two types of simultaneous error are conceivable: (1) detecting a region of significance when in fact there is none; and (2) the region of significance contains the point for which two populations are equal in expected criterion score. The former type of error was investigated by Shields (1978). The rate of this type of error associated with the J-N technique was found to be approximately .15 at a nominal α of .05 in a complete null data condition (two population regression lines are identical). The latter type of error under heterogeneous population slopes condition was explored by Chou and Huberty (1992). It was surprisingly found that the rate of this type of error associated with the original J-N

technique was approximately at the nominal α level. Based on these empirical results, it appears that the original J-N technique can be used to make simultaneous inferences provided that the error rate of the first type can be controlled. The overall null hypothesis test of no region of significance as presented in the original paper of Johnson and Neyman (1936) might be needed to serve this purpose.

Without performing the first step (omnibus hypothesis testing procedure), educational researchers typically go straight to the computation of R. Rogosa (1980) has shown through some algebraic manipulations that R is composed of all covariate points (denote the covariate by X) which satisfy the second-degree inequality of the form, $AX^2 + 2BX + C > 0$. The left hand side of the inequality takes on the form of a parabola. When setting it to 0, we get two bounds for R, denoted as X_- (for the smaller root) and X_+ (for the larger root). It is important to note that the sign of A determines the form of R in terms of X_- and X_+. The inequality has no real solutions when $B^2 - AC$ is negative. A common practice of using the J-N technique is when the test of slopes homogeneity is rejected. It can be shown that the sign of A is positive when the slope homogeneity is rejected. Under such a situation the parabola opens upward, and consequently the region of significance will always exist. Therefore, the J-N technique has been presented as to yield a region of non-significance between X_- and X_+ (Huitema, 1980; Pedhazur, 1982). According the empirical results due to Chou and Huberty (1992), it appears that simultaneous inferences may be appropriate for R, computed after the rejection of the slope homogeneity assumption. However, a region of significance also exists when two regression lines are parallel and non-null (i.e., equal slopes, different intercepts). Rogosa (1980) argued that the J-N technique could be used regardless of the assumption of slope homogeneity being rejected or not. When the test of slope homogeneity is not rejected (A is negative), the parabola opens downward. Consequently, R is composed of the X points between X_- and X_+. This paper is to examine the appropriateness of making simultaneous inferences for R or R' under two situations: (1) when the test of the slope homogeneity assumption is rejected, and (2) when the test of the slope homogeneity assumption is not rejected. The necessity of the omnibus null hypothesis testing step for the original J-N technique is examined under each situation.

MONTE CARLO SIMULATION PROCEDURES

Three settings of regression coefficients were selected in computer Monte Carlo simulations. They are shown in Table 1. The variances of the covariate and the random error component were set as 9 and 36, respectively. The outcome variable under investigation was the simultaneous error of the second type. In each setting, the regression parameters were determined such that the two population regression lines intersect in the middle of the covariate data range (grand covariate mean X=20). The three settings differ in the extent of slopes heterogeneity (slight, moderate, and severe in the author's arbitrary judgement). Within each setting, three sample size combinations were considered (10:10, 20:20, and 30:30) with 10,000 simulated experiments for each sample size combination. The J-N technique was applied in each simulated experiment. With nine artificially generated data conditions (three regression settings and three sample size combinations), the total number of simulated experiments in this study was 90,000.

SIMULATION RESULTS

The performance of the omnibus test in the settings of slight, moderate, and severe heterogeneous slopes was shown in Table 2 through Table 4 under the "Omnibus Test" heading. The total number of rejection of the omnibus hypothesis, the number of incorrect R's computed following the rejection of the omnibus hypothesis, and the percent of the incorrect R's were reported under this heading. The omnibus test obviously failed to control the proportion of the incorrect R's at the nominal alpha level in the slight heterogeneous slopes setting. As the severity of slope heterogeneity increased, the total percentage of incorrect R's approached the nominal value of .05 (see Table 3 and Table 4). There was a tendency of the error rate dropping as the sample size increased. The $\bar{X}_-$ and $\bar{X}_+$ reported were the two average lower and upper boundaries of the non-significant region. Note that $\bar{X}_-$ and $\bar{X}_+$ were not reported for $A < 0$ because the computed R would make little sense under the current simulated settings.

Table 2 through Table 4 also showed that the computation of R for $A < 0$ after a significant omnibus null hypothesis test produced incredibly high simultaneous error rates. On the other hand, for $A > 0$, the empirical simultaneous error rates got quite close to the nominal α (the largest error rate was .085, found in Table 2 at sample size 10:10).

Table 5 though Table 7 reported the simulation results without performing an omnibus test. The importance of the slope homogeneity test was strongly revealed. For $A < 0$, simultaneous error rates were unacceptbly high, whereas for $A > 0$, the simultaneous error rates generally were controlled at the nominal alpha level. For $A > 0$, the simultaneous error rate was unacceptably large only when sample sizes were small, found in Table 5 at sample size 10:10.

DISCUSSION

The omnibus hypothesis testing procedure as proposed in the original paper of Johnson and Neyman (1936) was unable to control the simultaneous error rate at the nominal α level. The first step of the J-N technique appeared to be unnecessary. Because the covariates used in most educational studies are often random, making non-simultaneous inference at any single covariate point in R is seldomly useful. Fortunately, the original J-N technique appeared to be still be valid for making simultaneous inferences over the entire range of R, given that the test of slope homogeneity assumption is rejected. A warning should be made against the use of the J-N technique when the test of slope homogeneity is not rejected. In the light of the results from this study, the common practice for using the J-N technique only when the test of slope homogeneity assumption is rejected is still recommended. Another finding of this study worth attention is that sufficient sample sizes must be obtained for the application of the J-N technique. In the settings of current study, sample sizes larger than 20:20 were required to secure the validity of the simultaneous inference.

Nevertheless, the ability of the original J-N technique to control the simultaneous error rate at the nominal level should not be overstated. In comparing two regression lines, a region of significance will always exist if two lines are not identical and sample sizes are sufficiently large. In fact, when the two different population regression lines are compared, the two population means are equal only at the covariate point where the two lines intersect. Some of the differences between two

population regression lines may be trivial to be declared as practically significant. Cohen (1988) stressed the need for an awareness of the effect magnitude over and above which the researcher may wish to conclude that the difference between two population means are practically significant. With this idea in mind, future users of the J-N technique may want to find a region of significance which would allow a researcher-determined magnitude of significant difference. This may warrant further studies.

REFERENCES

Bickel, P. J., & Doksum, K. A. (1977). Mathematical statistics: Basic ideas and selected topics. San Francisco: Holden-Day

Chou, T., & Huberty, C. J (April, 1992). The robustness of the Johnson-Neyman Technique, Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Cohen, J. (1988). Statistical Power Analysis for Behavioral Sciences New York: Academic Press.

Huitema, B. E. (1980). The Analysis of Covariance and Alternatives, New York: Wiley

Johnson, P. O., & Neyman, J. (1936). Tests of certain linear hypotheses and their applications to some educational problems. Statistical Research Memoirs, 1, 57-93

Pedhazur, E. J. (1982). Multiple Regression in Behavioral Research. New York: Holt, Rinehart, & Winston.

Potthoff, R. F. (1964). On the Johnson-Neyman technique and some extensions thereof. Psychometrika, 29, 241-256.

Rogosa, D. (1980). Comparing nonparallel regression lines. Psychological Bulletin, 88, 307-321.

Shields, J. L. (1978). An investigation of the effect of heteroscedasticity and heterogeneity of variance on the analysis of covariance and the Johnson-Neyman technique. Technical Paper 292, U.S. Army Research Institute for the Behavioral and Social Sciences, Alexandria, Virginia

Table 1

Regression Coefficients in the Three Simulation Settings

Values of (a_0, a_1) for Group 1 and (b_0, b_1) for Group 2

Extent of Slope Heterogeneity	Small	Medium	Large
Group 1	40, .5	30, 1	20, 2
Group 2	60,-.5	70,-1	100,-2

Table 2

Simulation Results Under Slight Extent of Slope Heterogeneity

R Obtained After a Significant Omnibus Test

Size	Omnibus Test		Slope Homegeneity Test	Incorr. n	R	P	$\bar{X}_1$	$\bar{X}_2$
10:10	Total Rejections	1259	A < 0	288	214	.74	•	•
	Total Incorr. R's	340						
	Percent	.27	A > 0	971	126	.13	13.27	26.74
20:20	Total Rejections	2414	A < 0	271	185	.68	•	•
	Total Incorr. R's	370						
	Percent	.16	A > 0	2143	185	.09	14.07	25.87
30:30	Total Rejections	3622	A < 0	269	180	.67	•	•
	Total Incorr. R's	398						
	Percent	.11	A > 0	3353	218	.07	14.07	25.72

The slope combination is (.5, -.5).

Table 3

Simulation Results Under Moderate Extent of Slope Heterogeneity

R Obtained After a Significant Omnibus Test

Size	Omnibus Test		Slope Homogeneity Test	Incorr. n	R	P	$\bar{X}_.$	$\bar{X}_{.'}$
10:10	Total Rejections	3644	A < 0	303	150	.50	•	•
	Total Incorr. R's	433						
	Percent	.12	A > 0	3341	283	.09	14.51	24.51
20:20	Total Rejections	7261	A < 0	144	64	.44	•	•
	Total Incorr. R's	469						
	Percent	.06	A > 0	7117	405	.06	16.54	23.40
30:30	Total Rejections	8986	A < 0	50	23	.46	•	•
	Total Incorr. R's	524						
	Percent	.06	A > 0	8936	501	.06	17.56	22.42

The slope combination is (1, -1).

Table 4

Simulation Results Under Severe Extent of Slope Heterogeneity

R Obtained After a Significant Omnibus Test

Size	Omnibus Test		Slope Homogeneity Test	Incorr. n	R	P	$\bar{X}_-$	$\bar{X}_+$
10:10	Total Rejections	8845	A < 0	94	27	.29	•	•
	Total Incorr. R's	527						
	Percent	.06	A > 0	8751	500	.057	17.52	22.37
20:20	Total Rejections	9975	A < 0	0	0	0	•	•
	Total Incorr. R's	467						
	Percent	.05	A > 0	9975	467	.047	18.85	21.14
30:30	Total Rejections	10000	A < 0	0	0	0	•	•
	Total Incorr. R's	491						
	Percent	.049	A > 0	10000	491	.049	19.15	20.85

*

The slope combination is (2, -2).

Table 5

Simulation Results Under Slight Extent of Slope Heterogeneity

R Computed Without An Omnibus Test

Size	Total Obtainable R	Slope Homogeneity Test	Incorr. n	R	P	$\bar{X}_.$	$\bar{X}_.$
10:10	2603	A < 0	1069	397	.37	.	.
		A > 0	1534	130	.12	7.96	35.25
20:20	4236	A < 0	1130	297	.263	.	.
		A > 0	3106	170	.055	-10.31	32.88
30:30	5647	A < 0	1122	265	.236	.	.
		A > 0	4525	226	.050	5.11	30.77

Table 6

Simulation Results Under Moderate Extent of Slope Heterogeneity

R Computed Without An Omnibus Test

Size	Total Obtainable R	Slope Homogeneity Test	Incorr. n	R	P	$\bar{X}_.$	$\bar{X}_.$
10:10	5724	A < 0	1141	220	.193	.	.
		A > 0	4583	307	.067	4.64	35.45
20:20	8645	A < 0	546	77	.141	.	.
		A > 0	8099	390	.048	14.20	30.09
30:30	9639	A < 0	184	31	.168	.	.
		A > 0	9455	460	.049	16.16	23.62

Table 7

Simulation Results Under Severe Extent of Slope Heterogeneity

R Computed Without An Omnibus Test

Size	Total Obtainable R	Slope Homogeneity Test	Incorr. n	R	P	$\bar{X}_-$	$\bar{X}_+$
10:10	9538	A < 0	268	27	.101	.	.
		A > 0	9270	500	.054	16.49	22.99
20:20	9995	A < 0	1	0	0	.	.
		A > 0	9994	467	.047	18.84	21.15
30:30	10000	A < 0	0	0	0	.	.
		A > 0	10000	491	.049	19.15	20.85